PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-149370
(43)Date of publication of application : 02.06.1998

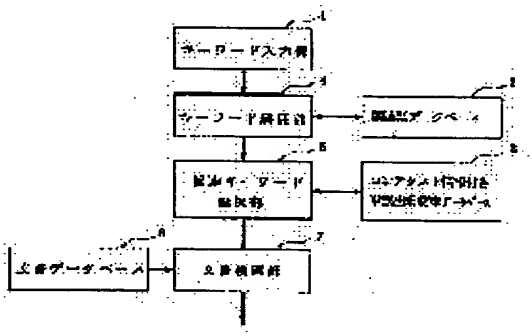
(51)Int.Cl. G06F 17/30
G06F 17/21

(21)Application number : 08-320828
(22)Date of filing : 15.11.1996

(71)Applicant : NEC CORP
(72)Inventor : OKUMURA AKITOSHI

(54) DOCUMENT RETRIEVAL METHOD AND DEVICE USING CONTEXT INFORMATION

(57)Abstract:
PROBLEM TO BE SOLVED: To improve the document retrieval accuracy by selecting a word having more intensive relationship as a keyword in a semantic sense when the relative words are added as keywords.
SOLUTION: A keyword expansion part 4 extracts the relative words of an input keyword from a relative word data base, and a relative keyword selection part 5 outputs all relative keyword candidates expanded to the keyword in sequence and in a processing symmetric form. The part 5 also checks whether the relative keyword candidates have the same context information as the input keyword based on a word appearance frequency data base 3 including the context information. Thus, the part 5 selects the relative keywords having high appearance frequency. Then a document retrieval part 7 retrieves a document from a document data base 6 and outputs it based on the keyword that is inputted to a keyword input part 1 and the keyword that is selected at the part 5.



LEGAL STATUS

[Date of request for examination]	15.11.1996
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	2894301
[Date of registration]	05.03.1999
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

Ref. 1

98149370 (1653x2338x2 tiff)

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－149370

(43)公開日　平成10年(1998)６月２日

(51)Int.Cl.[6]		識別記号	ＦＩ		
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/403	３２０Ｄ
	17/21			15/20	５９０Ｅ
				15/40	３７０Ａ
				15/403	３４０Ｂ
					３５０Ｃ

審査請求　有　請求項の数９　ＦＤ　（全　９　頁）

(21)出願番号　特願平8－320828

(22)出願日　平成８年(1996)11月15日

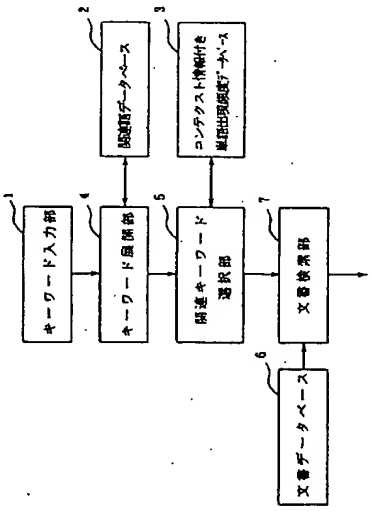
(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　奥村　明俊
東京都港区芝五丁目７番１号　日本電気株式会社内

(74)代理人　弁理士　加藤　朝道

(54)【発明の名称】　文脈情報を用いた文書検索方法および装置

(57)【要約】
【課題】関連性のある単語をキーワードとして追加する場合に、より意味的に関連性の強い単語を選択してキーワードとして加え、検索精度を向上する、文書検索方法及び装置の提供。
【解決手段】キーワードを入力するキーワード入力部１と、キーワード入力部１より入力された入力キーワードから、類語辞書、関連語辞書、シソーラス辞書などを用いて同義・類義語など関連単語を展開するキーワード展開部４と、展開された関連単語から、単語の共起関係と文脈情報と頻度を保持した単語共起データベース３を参照して、入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択する関連キーワード選択部５と、入力キーワードおよび前記関連キーワードを検索キーワードとして、文書データベース６から文書の検索を行なう文書検索部７と、を備える。



1

【特許請求の範囲】

【請求項１】入力されたキーワードから検索キーワードを拡張する場合に、類語辞書、関連語辞書、シソーラス辞書などを用いて関連単語を展開し、

単語の共起関係と文脈情報と頻度を保持した単語共起データベースを用いて、前記入力されたキーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択し、検索キーワードに追加して、文書を検索する、ことを特徴とする文書検索方法。

【請求項２】前記検索キーワードを拡張する場合、出現頻度が予め定められた所定の値よりも少ない特徴的な入力キーワードについてのみ、関連単語の展開を行ない、検索キーワードに追加して、文書を検索する、ことを特徴とする請求項１記載の文書検索方法。

【請求項３】入力されたキーワードから検索キーワードを拡張する場合に、

類語辞書、関連語辞書、シソーラス辞書などを用いて関連単語を展開し、

単語の共起関係と文脈情報と頻度を保持した単語共起データベースを用いて、前記入力されたキーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択し、

さらに、前記関連キーワードから再帰的に関連単語の展開と、関連キーワード選択を行ない、

検索キーワードに追加して、文書を検索する、ことを特徴とする文書検索方法。

【請求項４】前記検索キーワードを拡張する場合、出現頻度が予め定められた所定の値よりも少ない特徴的な関連キーワードについてのみ、再帰的な関連単語の展開と関連キーワードの選択を行ない、検索キーワードに追加して、文書を検索することを特徴とする請求項３記載の文書検索方法。

【請求項５】キーワードを入力するキーワード入力部と、

前記キーワード入力部より入力された入力キーワードから、類語辞書、関連語辞書、シソーラス辞書などを用いて同義・類義語など関連単語を展開するキーワード展開部と、

展開された関連単語から、単語の共起関係と文脈情報と頻度を保持した単語共起データベースを参照して、前記入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択する関連キーワード選択部と、

前記入力キーワードおよび前記関連キーワードを検索キーワードとして、文書データベースから文書の検索を行なう文書検索部と、

を備えたことを特徴とする文書検索装置。

【請求項６】キーワードを入力するキーワード入力部と、

前記キーワード入力部より入力された入力キーワードから、類語辞書、関連語辞書、シソーラス辞書などを用いて同義・類義語など関連単語を展開する際、出現頻度が少ない特徴的な入力キーワードの場合のみ関連単語の展開を行なう選択的キーワード展開部と、

展開された関連単語から、単語の共起関係と文脈情報と頻度を保持した単語共起データベースを参照して、入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択する関連キーワード選択部と、

前記入力キーワードおよび前記関連キーワードを検索キーワードとして、文書データベースから文書の検索を行なう文書検索部と、

を備えたことを特徴とする文書検索装置。

【請求項７】キーワードを入力するキーワード入力部と、

前記キーワード入力部より入力された入力キーワードと関連キーワードから類語辞書、関連語辞書、シソーラス辞書などを用いて同義・類義語および関連単語を展開するキーワード展開部と、

展開された関連単語から、単語の共起関係と文脈情報と頻度を保持した単語共起データベースを参照して、前記入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択し、前記キーワード展開部に出力する再帰的関連キーワード選択部と、

前記入力キーワードおよび前記関連キーワードを検索キーワードとして文書データベースから文書の検索を行なう文書検索部と、

を備えたことを特徴とする文書検索装置。

【請求項８】前記関連キーワードが、出現頻度が予め定められた所定の値よりも少ない特徴的な関連キーワードである場合のみ、キーワード展開を行なう、ことを特徴とする請求項７記載の文書検索装置。

【請求項９】前記選択的キーワード展開部が、前記入力キーワードの出現頻度を予め定められた閾値と比較し、出現頻度が前記閾値を越えない入力キーワードについてのみ、関連単語を展開して、前記関連キーワード選択部に出力する、ことを特徴とする請求項５記載の文書検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書検索方法及び装置に関し、特に単語共起情報を用いて入力キーワードを拡張して文書を検索する方法および装置に関する。

【０００２】

【従来の技術】従来、この種の文書検索方法は、文書検索装置などで、入力された文書やキーワードから検索文字列を設定し、その検索文字列を含む文書を検索するために用いられている。

【０００３】従来の文書検索方法の一例として、例えば特開平３－１７２９６６号公報には、フルテキストのデ

ータベースの中から類似文書を効率よく、かつ高精度に検索ができる類似文書検索装置の構成が提案されている。

【0004】この従来の類似文書検索装置は、文書を直接入力し、コード情報とする構文書入力部と、入力された文字列を分かち書きし形態素情報を付与するとともに、形態素情報を基にして文書（文節間）の係り受け構造を判定する係り受け解析部と、この係り受け解析結果から文構造を決定し、この文構造から索引を抽出するとともに索引の重要度を付与する索引抽出部と、入力文書、係り受け解析結果、索引抽出結果を蓄積する文書蓄積部と、前記索引抽出部の索引をシソーラス辞書で展開するシソーラス展開部と、入力文書と蓄積されている文書との類似度を索引の類似度と係り受け関係の類似度から判定する類似文書検索部と、検索した類似文書を出力する類似文書出力部と、を備えて構成されている。

【0005】シソーラス展開部では、多義判定テーブルが用意されており、表記上は同じでも意味が異なる単語の区別を、文書の分野によって判定する。すなわち、多義テーブルは、単語の表記、読み、利用分野の情報からなり、シソーラス展開する場合、入力された文書の分野に最も意味的に正しい同義語、類義語を出力する。例えば、「CD」といった場合、銀行関係の分野では「キャッシュ・ディスペンサー」、音楽関係の分野では「コンパクト・ディスク」という具合に、その分野に対応する同義語、類義語を出力する。

【0006】

【発明が解決しようとする課題】しかしながら、上記した従来の文書検索装置においては、多義判定テーブルによるシソーラス展開では、展開された関連語彙の優先度を判定することができないので、必ずしも適当な関連語彙を選択できず、検索精度が向上しない、という問題点を有している。

【0007】その理由は、多義判定テーブルでは、文書の分野情報を特定して同じ分野の語彙を選択するものであるが、文書に分野情報が記述されていない場合には、どの語彙を関連語彙とするか、判定することができない、ためである。

【0008】したがって、本発明は、上記問題点に鑑みてなされたものであって、その目的は、関連性のある単語をキーワードとして追加する場合に、より意味的に関連性の強い単語を選択してキーワードとして加えて検索することにより、検索精度を向上する文書検索方法及び装置を提供することにある。

【0009】

【課題を解決するための手段】前記目的を達成するため、本発明の第１の文書検索方法（請求項１）は、入力されたキーワードから検索キーワードを拡張する場合に、類語辞書、関連語辞書、シソーラス辞書などを用いて関連単語を展開し、単語の共起関係と文脈情報と頻度を保持した単語共起データベースを用いて、入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択し、検索キーワードに追加して文書検索を行なうことを特徴とする。

【0010】本発明の第２の文書検索方法（請求項２）は、前記第１の文書検索方法において、検索キーワードを拡張する場合、出現頻度が予め定められた所定の値よりも少ない特徴的な入力キーワードについてのみ関連単語の展開を行ない、検索キーワードに追加して検索することを特徴とする。

【0011】本発明の第３の文書検索方法（請求項３）は、入力されたキーワードから検索キーワードを拡張する場合に、類語辞書、関連語辞書、シソーラス辞書などを用いて関連単語を展開し、単語の共起関係と文脈情報と頻度を保持した単語共起データベースを用いて、入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択し、さらに関連キーワードから再帰的に関連単語の展開と関連キーワード選択を行ない検索キーワードに追加して文書検索を行なうことを特徴とする。

【0012】本発明の第４の文書検索方法（請求項４）は、前記第３の文書検索方法において、検索キーワードを拡張する場合、出現頻度が予め定められた所定の値よりも少ない特徴的な関連キーワードについてのみ再帰的な関連単語の展開と関連キーワードの選択を行ない、検索キーワードに追加して検索することを特徴とする。

【0013】本発明の第１の文書検索装置（請求項５）は、キーワードを入力するキーワード入力部と、このキーワード入力部より入力された入力キーワードから類語辞書、関連語辞書、シソーラス辞書などを用いて同義・類義語など関連単語を展開するキーワード展開部と、展開された関連単語から、単語の共起関係と文脈情報と頻度を保持した単語共起データベースを参照して、入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択する関連キーワード選択部と、前記入力キーワードおよび前記関連キーワードを検索キーワードとして文書データベースから文書の検索を行なう文書検索部とを有することを特徴とする。

【0014】本発明の第２の文書検索装置（請求項６）は、前記第１の文書検索装置のキーワード展開部を選択的キーワード展開部によって置き換え、出現頻度が少ない特徴的な入力キーワードについてのみ関連単語の展開を行なう検索キーワードに追加して検索することを特徴とする。

【0015】本発明の第３の文書検索装置（請求項７）は、キーワードを入力するキーワード入力部と、このキーワード入力部より入力された入力キーワードと関連キーワードから類語辞書、関連語辞書、シソーラス辞書などを用いて同義・類義語および関連単語を展開するキーワード展開部と、展開された関連単語から、単語の共起

5

関係と文脈情報と頻度を保持した単語共起データベースを参照して、入力キーワードと同じ文脈情報をもつ共起頻度の高い関連単語を関連キーワードとして選択しキーワード展開部に出力する再帰的関連キーワード選択部と、前記入力キーワードおよび前記関連キーワードを検索キーワードとして文書データベースから文書の検索を行なう文書検索部とを有することを特徴とする。

【００１６】本発明の第４の文書検索装置（請求項８）は、前記第３の文書検索装置のキーワード展開部において、出現頻度が予め定められた所定の値よりも少ない特徴的な関連キーワードについてのみ再帰的な関連単語の展開を行ない、検索キーワードに追加して検索することを特徴とする。

【００１７】

【発明の実施の形態】次に、本発明の実施の形態について図面を参照して詳細に説明する。

【００１８】図１は、本発明の第１の実施の形態に係る文書検索装置の構成を示すブロック図である。図１を参照すると、本発明の第１の実施の形態に係る文書検索装置は、キーワード入力部１と、コンテクスト情報付き単語出現頻度データベース３と、関連語データベース２、キーワード展開部４と、関連キーワード選択部５と、文書データベース６と、文書検索部７と、を含んで構成されている。

【００１９】キーワード入力部１は、キーボード等によって単数または複数のキーワードを検索文字列として入力する。

【００２０】コンテクスト情報付き単語出現頻度データベース３は、文書データベース６もしくはその他の文書データベースに出現する単語が、他のどのような単語とどの程度の頻度で出現したかを単語の左右両側のコンテクスト情報とともに記述したデータベースである（単語の共起関係と文脈情報と頻度を保持した単語共起データベース）。

【００２１】コンテクスト情報の一例は、ある単語の左右に存在する単語が、助詞の場合にはその文字列、助詞以外の場合にはその品詞からなり、「単語（左コンテクスト、右コンテクスト）、頻度」のような形式で表現され、出現頻度は、大中小の３段階で記録する（図３参照）。この「大」、「中」、「小」の類別は頻度の数値データであってもよいことは勿論である。

【００２２】図３は、コンテクスト情報付き単語出現頻度データベース３のうち人員削減の内容を例示したものである。例えば、「人員削減（名の、によって）、大」（図３の２行目）は、単語「人員削減」が、左コンテクスト「名の」、及び右コンテクスト「によって」と共に出現する（すなわち「…名の人員削減によって…」）頻度が「大」であることを示し、また「人員削減（の、名詞）、中」（図３の５行目）は、「の人員削減（名詞）」（例えば「…の人員削減計画…」）の出現頻度が

6

「中」であることを示している。

【００２３】関連語データベース２は、シソーラス辞書や、類義語辞書、関連語辞書である。

【００２４】キーワード展開部４は、キーワード入力部１から与えられた単数または複数の入力キーワードから関連語データベース２を用いて、関連単語を関連キーワード候補として、関連キーワード選択部５に出力する。

【００２５】関連キーワード選択部５は、コンテクスト情報付き単語出現頻度データベース３を用いて、関連キーワード候補の中から、入力キーワードと同じコンテクスト情報をもつ高頻度の単語を関連キーワードとして選択する。

【００２６】図２は、本発明の実施の形態における、関連キーワード選択部５の処理フローを説明するためのフローチャートである。

【００２７】図２を参照すると、関連キーワード選択部５は、関連キーワード候補を出力するステップＡ１と、この関連キーワード候補が入力キーワードと同じコンテクストで出現するかを判定するステップＡ２（「文脈共起関係判定ステップ」という）と、出現頻度が高い場合には、該候補を関連キーワードとして選択する関連キーワード選択ステップＡ３と、関連キーワード候補が他にもあるか否かを判定するステップＡ４と、からなる。

【００２８】文書データベース６は、電子化された文書を格納しているデータベースである。

【００２９】文書検索部７は、キーワード入力部１に入力されたキーワードと関連キーワード選択部５で選択された関連キーワードとを用いて、文書データベース６より文書を検索し出力する。

【００３０】次に、このように構成された第１の実施の形態に係る文書検索装置の動作について、図１、図２、図３および図４を参照して説明する。図４は、関連語データベース２の内容を例示したものと、それぞれの関連語について、コンテクスト情報付き単語出現頻度データベース３から出現頻度の大きい文脈情報を抽出した内容を例示したものである。

【００３１】キーワード入力部１から入力キーワードが入力されると、キーワード展開部４に供給される。

【００３２】キーワード展開部４は、関連語データベースから入力キーワードの関連単語を抽出する。

【００３３】関連キーワード選択部５は、キーワード展開されたすべての関連キーワード候補を順に処理対象として出力する（図２のステップＡ１）。

【００３４】次に、関連キーワード選択部５は、コンテクスト情報付き単語出現頻度データベース３を用いて、関連キーワード候補が、入力キーワードと同じコンテクスト情報をもつ単語かどうかを調べる（図２のステップＡ２）。ステップＡ２の判定の結果、出現頻度が高い場合、関連キーワードとして選択する（図２のステップＡ３）。続いて、関連キーワード展開部４は、他にも関連

キーワード候補があるかどうかを判定し（図2のステップA4）、残されていれば、ステップA1に制御を戻し、関連キーワード候補がなくなるまで、上記ステップA1～A4を繰り返す。

【0035】例えば図4を参照すると、キーワード展開部4は、「人員削減」の関連キーワード候補として、「大手航空会社」、「経営」、「合理化」、「人員整理」、「社員」、「希望退職」、「退職金」、「リストラ」を出力する。

【0036】図3を参照すると、関連キーワード選択部5は、コンテクスト情報付き単語出現頻度データベース3から、「人員削減」の高頻度文脈として、「人員削減（名の、を）」、「人員削減（名の、によって）」、及び「人員削減（名の、に対する）」を抽出する。

【0037】これらの頻度が大きいのは、例えば、「…名の人員削減を…」、「…名の人員削減に対する…」、「…名の人員削減によって…」、という表現が一般によく使われることを示している。

【0038】関連キーワード選択部5は、キーワード展開部4で出力された候補（上記した「大手航空会社」、「経営」、「合理化」、「人員整理」、等）に対して、コンテクスト情報付き単語出現頻度データベース3から頻度の大きい文脈を出力する。

【0039】続いて、関連キーワード選択部5は、関連キーワード候補の中から「人員削減」と同じ文脈の頻度が大きい、「人員整理」、「希望退職」、「リストラ」を関連キーワードとして選択する。

【0040】文書検索部7は、キーワード入力部1から入力された入力キーワードと、関連キーワード選択部5で選択された関連キーワードと、を用いて、文書データベース6より、文書を検索して、出力する。

【0041】このように、本発明の第1の実施の形態においては、入力キーワードから関連語データベース2を用いて関連単語を展開し、関連単語からコンテクスト情報付き単語出現頻度データベース3を用いて、入力キーワードと同じ文脈で出現する関連単語を選択して、文書の検索に用いることができる。

【0042】また、本発明の第1の実施の形態においては、関連単語が複数存在する場合でも、文脈情報と共起関係を用いることによって、関連性の低い単語を排除することができる。このため、文書に分野情報が記述されていない場合でも、文書の検索精度を格段に向上する。

【0043】図5は、本発明の第2の実施の形態に係る文書検索装置の構成を示すブロック図である。

【0044】図5を参照すると、本発明の第2の実施の形態に係る文書検索装置においては、図1に示した前記第1の実施の形態に係る文書検索装置のキーワード展開部4が、選択的キーワード展開部4′で置き換えられている点が相違しており、その他の構成は同様とされている。

【0045】図6は、本発明の第2の実施の形態における選択的キーワード展開部4′の処理フローを説明するためのフローチャートである。図6を参照すると、選択的キーワード展開部4′は、入力キーワードを出力するステップB1と、入力キーワードの出現頻度は少ないか否かを判定するステップB2（「展開判定ステップ」という）、関連単語を展開し関連キーワード選択部5へ出力するステップB3（「キーワード展開ステップ」という）と、入力キーワードが他にも有るか否かを判定するステップB4からなる。

【0046】すなわち、選択的キーワード展開部4′において、ステップB1で入力キーワードを出力し、ステップB2において出現頻度が、設定された閾値を越えない場合、ステップB3にて、関連単語を展開し、関連キーワード選択部5に出力する処理を行う。そして、入力キーワードが残っている場合には、同様にステップB1から行なう。

【0047】このように、本発明の第2の実施の形態においては、選択的キーワード展開部4′によって、出現頻度が高い極めて一般的な単語に関するキーワード展開を抑制することができ、キーワードが増え過ぎることによる検索精度の低下を、抑制することができる。

【0048】図7は、本発明の第3の実施の形態に係る文書検索装置の構成を示すブロック図である。

【0049】図7を参照すると、本発明の第3の実施の形態に係る文書検索装置においては、図1に示した前記第1の実施の形態に係る文書検索装置の関連キーワード選択部5が再帰的関連キーワード選択部5′で置き換えられている点と、再帰的関連キーワード選択部5′の出力がキーワード展開部4にも出力される点が相違しており、その他の構成は同様とされている。

【0050】再帰的関連キーワード選択部5′は、コンテクスト情報付き単語出現頻度データベース3を用いて、関連キーワード候補の中から、入力キーワードと同じコンテクスト情報をもつ高頻度の単語を関連キーワードとして選択するとともに、キーワード展開部4へ関連キーワードを出力する。

【0051】キーワード展開部4に送られた関連キーワードは、入力キーワードと同じように、関連語データベース2によって関連単語が抽出され、再帰的関連キーワード選択部5′によって、関連キーワードの関連キーワードが選択される。

【0052】この処理は予め与えられた数のキーワードが得られるまで、繰り返される。

【0053】例えば、図4を参照すると、キーワード展開部4は、「人員削減」の関連キーワード候補として、「大手航空会社」、「経営」、「合理化」、「人員整理」、「社員」、「希望退職」、「退職金」、「リストラ」を出力する。

【0054】図3を参照すると、再帰的関連キーワード

選択部5′は、コンテクスト情報付き単語出現頻度データベース3から、「人員削減」の高頻度文脈として、「人員削減（名の、を）」、「人員削減（名の、によって）」、及び「人員削減（名の、に対する）」を抽出する。

【0055】再帰的関連キーワード選択部5′は、キーワード展開部4で出力された候補（上記した「大手航空会社」、「経営」、「合理化」、「人員整理」、等）に対して、コンテクスト情報付き単語出現頻度データベース3から頻度の大きい文脈を出力する。

【0056】続いて、再帰的関連キーワード選択部5′は、関連キーワード候補の中から「人員削減」と同じ文脈の類度が大きい、「リストラ」、「人員整理」、「希望退職」を関連キーワードとして選択するとともに、キーワード展開部4に出力する。

【0057】図9は、関連語データベース2の内容を例示したものと、それぞれの関連語についてコンテクスト情報付き単語出現頻度データベース3から出現頻度の大きい文脈情報を抽出した内容を例示したものである。

【0058】図9を参照すると、キーワード展開部4は、「リストラ」の関連キーワード候補として、「レイオフ」、「統廃合」、「再構築」を出力する。

【0059】図3を参照すると、再帰的関連キーワード選択部5′は、コンテクスト情報付き単語出現頻度データベース3から、「レイオフ」の高頻度文脈として、「レイオフ（名の、によって）」を抽出する。

【0060】再び図9を参照すると、再帰的関連キーワード選択部5′は、キーワード展開部4で出力された候補に対して、コンテクスト情報付き単語出現頻度データベース3から頻度の大きい文脈を出力する。

【0061】続いて、再帰的関連キーワード選択部5′は、関連キーワード候補の中から、「リストラ」と同じ文脈の類度が大きい、「レイオフ」を関連キーワードとして選択する。

【0062】このように、本発明の第3の実施の形態においては、再帰的関連キーワード選択部5′によって、入力キーワードが極めて少ない場合にも、検索キーワードを十分に得ることができ、キーワードが少な過ぎることによる検索精度の低下を抑制することができる。

【0063】なお、本発明の第3の実施の形態の変形例として、再帰的関連キーワード選択部5′は、関連キーワードの数が発散することを防ぐために、設定された出現頻度の閾値を越えない単語のみを処理の対象とする、ようにしてもよい。

【0064】図8は、本発明の第3の実施の形態及びその変形例における再帰的関連キーワード選択部5′の処理フローを説明するためのフローチャートである。図8を参照すると、再帰的関連キーワード選択部5′は、関連キーワードを出力するステップC1と、該関連キーワードの出現頻度が少ないか否かを判定するステップC2（「展開判定ステップ」という）と、関連単語を展開し再帰的関連キーワード選択部5′へ出力するステップC3（「キーワード展開出力ステップ」という）と、関連キーワードが他にもあるか否かを判定するステップC4と、からなる。

【0065】ステップC1において、関連キーワードを出力し、ステップC2において出現頻度が設定された閾値を越えない場合、関連単語を展開し再帰的関連キーワード選択部5′に出力する処理を行う（ステップC3）。そして、関連キーワードが残っている場合、同様にステップC1から行なう。

【0066】このように、本発明の第3の実施の形態及びその変形例では、入力キーワードが極めて少ない場合にも検索キーワードを十分に得ることができ、また関連キーワードから出現頻度が高い極めて一般的な単語に関するキーワード展開を抑制することができ、キーワードが少な過ぎること、及び、キーワードが増え過ぎることによる検索精度の低下を抑制することができる。

【0067】

【発明の効果】以上説明したように、本発明によれば下記記載の効果を奏する。

【0068】本発明の第1の効果は、入力キーワードと関連性の強いキーワードが選択されるために、検索精度を向上することができる、ということである。この結果、本発明は、文書の検索精度を向上する。

【0069】その理由は、本発明においては、入力キーワードと同じ文脈情報をもつ出現頻度の高い単語が関連キーワードとして選択され、検索キーワードに加えられて検索するためである。

【0070】また、本発明の第2の効果として、選択的キーワード展開部によって、出現頻度が高い極めて一般的な単語に関するキーワード展開を抑制することができ、キーワードが増え過ぎることによる検索精度の低下を、抑制することができる、ということである。

【0071】さらに、本発明の第3の効果として、再帰的関連キーワード選択部によって、入力キーワードが極めて少ない場合にも、検索キーワードを十分に得ることができ、キーワードが少な過ぎることによる検索精度の低下を抑制することができる、ということである。

【図面の簡単な説明】

【図1】本発明の第1の実施の形態に係る文書検索装置の構成を示すブロック図である。

【図2】本発明の第1の実施の形態における関連キーワード選択の処理を示すフローチャートである。

【図3】本発明の第1の実施の形態におけるコンテクスト情報付き単語出現頻度データベースの内容の具体例を例示する図である。

【図4】本発明の第1の実施の形態を説明するための図であり、関連語データベースの内容を例示したものと、それぞれの関連語についてコンテクスト情報付き単語出

11

現頻度データベースから出現頻度の大きい文脈情報を抽出した内容を例示したものである。

【図5】本発明の第2の実施の形態に係る文書検索装置の構成を示すブロック図である。

【図6】本発明の第2の実施の形態における選択的キーワード展開部の処理を示すフローチャートである。

【図7】本発明の第3の実施の形態に係る文書検索装置の構成を示すブロック図である。

【図8】本発明の第3の実施の形態及び変形例における再帰的関連キーワード選択部の処理を示すフローチャートである。

【図9】本発明の第3の実施の形態を説明するための図であり、関連語データベースの内容を例示したものと、

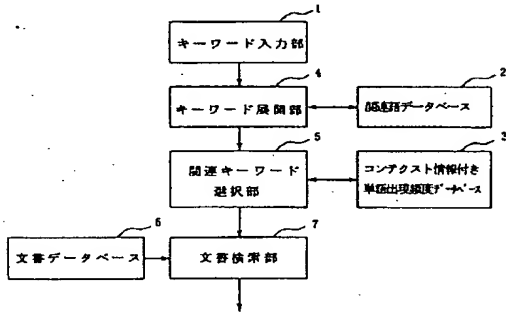
12

それぞれの関連語についてコンテクスト情報付き単語出現頻度データベースから出現頻度の大きい文脈情報を抽出した内容を例示したものである。

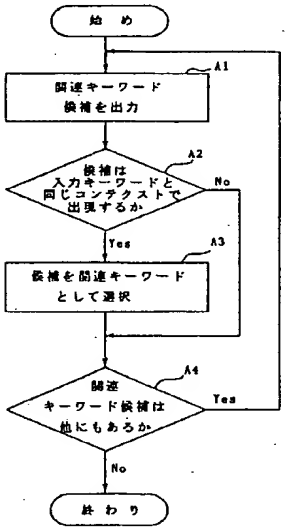
【符号の説明】

1　キーワード入力部
2　関連語データベース
3　コンテクスト情報付き単語出現頻度データベース
4　キーワード展開部
4′　選択的キーワード展開部
5　関連キーワード選択部
5′　再帰的関連キーワード選択部
6　文書データベース
7　文書検索部

【図1】



【図2】



【図3】

人員削減（左文脈、右文脈）	頻度
人員削減（名の、によって）	大
人員削減（名の、を）	大
人員削減（名の、に対して）	大
人員削減（の、名詞）	中
人員削減（区切り、名詞）	小
人員削減（した、名詞）	小
人員削減（名詞、名詞）	小
リストラ（名の、によって）	大
リストラ（の、名詞）	中

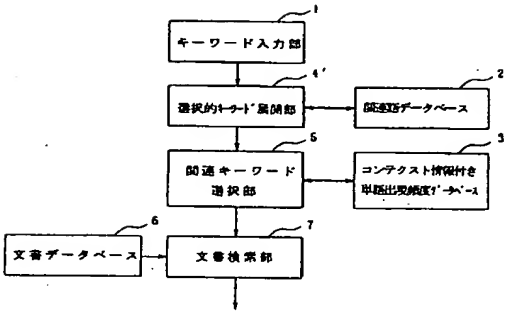
【図9】

	関連語データベース	頻度大の文脈
リストラ	レイオフ	（名の、によって）
	統廃合	（名詞、を）
	再構築	（名詞、を）

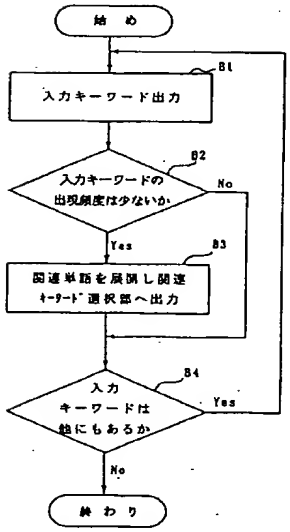
【図4】

	関連語データベース	頻度大の文節
人員削減	大手航空会社	（区切り、3社）
	経営	
	合理化	（経営、によって）
	人員整理	（名の、に対して）
	社員	
	希望退職	（名の、を）
	退職金	
	リストラ	（名の、によって）

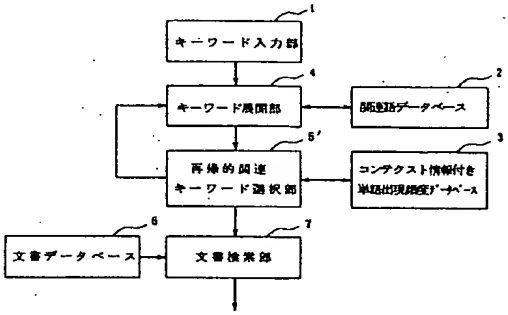
【図5】



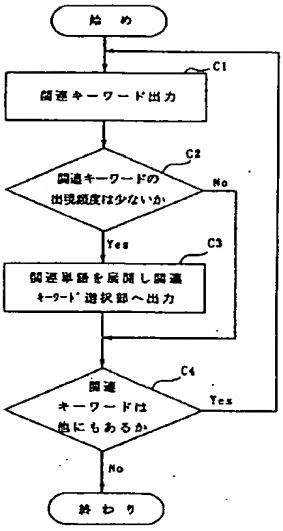
【図6】



【図7】



【図８】


始　め
C1
関連キーワード出力
C2
関連キーワードの
出現頻度は少ないか
No
Yes
C3
関連単語を反映し関連
ｷｰﾜｰﾄﾞ選択部へ出力
C4
関連
キーワードは
他にもあるか
Yes
No
終　わ　り